



Introduction to Bioinformatics

2. DNA Sequence Retrieval and comparison

Benjamin F. Matthews

United States Department of Agriculture

Soybean Genomics and Improvement

Laboratory

Beltsville, MD 20708

matthewb@ba.ars.usda.gov

What we will cover today

- ⌘ Retrieving a known DNA sequence
- ⌘ Similarity searching with a DNA sequence
- ⌘ BLAST

Retrieving a DNA sequence

You read a paper and..

- ★ Is full-length clone of gene available?
- ★ Is at least some of the DNA sequence available? (EST sequence)?

Finding Sequences in Databases

- ⌘ The public DNA and protein sequence databases are huge.
- ⌘ In order for these databases to be useful, the data must be readily accessible to researchers.

What Are You Looking For?

- ⌘ A gene?
 - DNA or protein sequence?
- ⌘ DNA sequences are essentially all in **GenBank**
 - Genomic, mRNA, cDNA, EST?
- ⌘ Proteins are harder to pin down
 - **GenPept** (GenBank Peptides) is huge and poorly annotated - lots of junk
 - **SwissProt** is carefully annotated, but not fully comprehensive
 - **PIR** is somewhere in between

Large Databases

- ⌘ Once upon a time, **GenBank** sent out sequence updates on CD-ROM disks a few times per year.
- ⌘ Now **GenBank** is over 95 Gigabytes (28 billion bases)
- ⌘ Most biocomputing sites update their copy of **GenBank** every day over the internet.
- ⌘ Scientists access **GenBank** directly over the Web

You can search DNA sequence database

- ⌘ Retrieve known sequences by
 - Keyword search
 - Accession numbers
- ⌘ If you know some DNA sequence
 - Compare your DNA sequence with those in database
 - Basic Local Alignment Search Tool (BLAST) searches

Retrieve a DNA sequence

- ✳ ENTREZ

- ◆ <http://www.ncbi.nlm.nih.gov/Entrez/>

- ✳ Click – Nucleotide

- ◆ GenBank

- ◆ OR

- <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>

- ✳ Accession number

- ✳ Keyword search

Entrez is a Tool for Finding Sequences

- ✳ **GenBank** is managed by the **NCBI** (National Center for Biotechnology Information) which is a part of the US National Library of Medicine.

- ✳ NCBI has created a Web-based tool called **Entrez** for finding sequences in **GenBank**.

- <http://www.ncbi.nlm.nih.gov>

- ✳ Each sequence in **GenBank** has a unique “**accession number**”.

- ✳ **Entrez** can also search for keywords such as gene names, protein names, and the names of organisms or biological functions

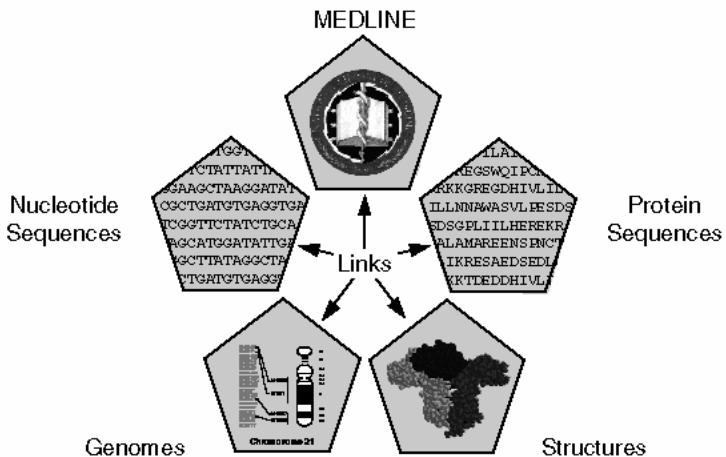
Entrez is a Database

- * The **Entrez** database contains all of the nucleotide and protein sequences in **GenBank** (updated daily) along with all of the literature in **MEDLINE** and the 3-D protein structures in **PDB (Protein Data Base)**.
- * **Entrez** is much more than a database, it is both a powerful search engine and a pre-computed list of relationships among all of its data elements

Entrez is Internally Cross-linked

- * DNA and protein sequences are linked to other similar sequences
- * **Medline** citations are linked to other citations that contain similar keywords
- * 3-D structures are linked to similar structures

Databases contain more than just DNA & protein sequences



NCBI Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | Entrez | Human Genome | GenBank | MapViewer | BLAST

Search across databases | Help

Welcome to the new Entrez cross-database search page

PubMed: biomedical literature citations and abstracts	Books: online books
PubMed Central: free, full text journal articles	ONIM: Online Mendelian Inheritance in Man
	Site Search: NCBI web and FTP sites

→

Nucleotide: sequence database (GenBank)	UniGene: gene-oriented clusters of transcript sequences
Protein: sequence database	CDD: conserved protein domain database
Genome: whole genome sequences	3D Domains: domains from Entrez Structure data
Structure: three-dimensional macromolecular structures	UniSTS: markers and mapping data
Taxonomy: organism in GenBank	PopSet: population study data sets
SNP: single nucleotide polymorphism	GEO: expression and molecular abundance profiles
Gene: gene-centered information	GEO DataSets: experimental sets of GEO data

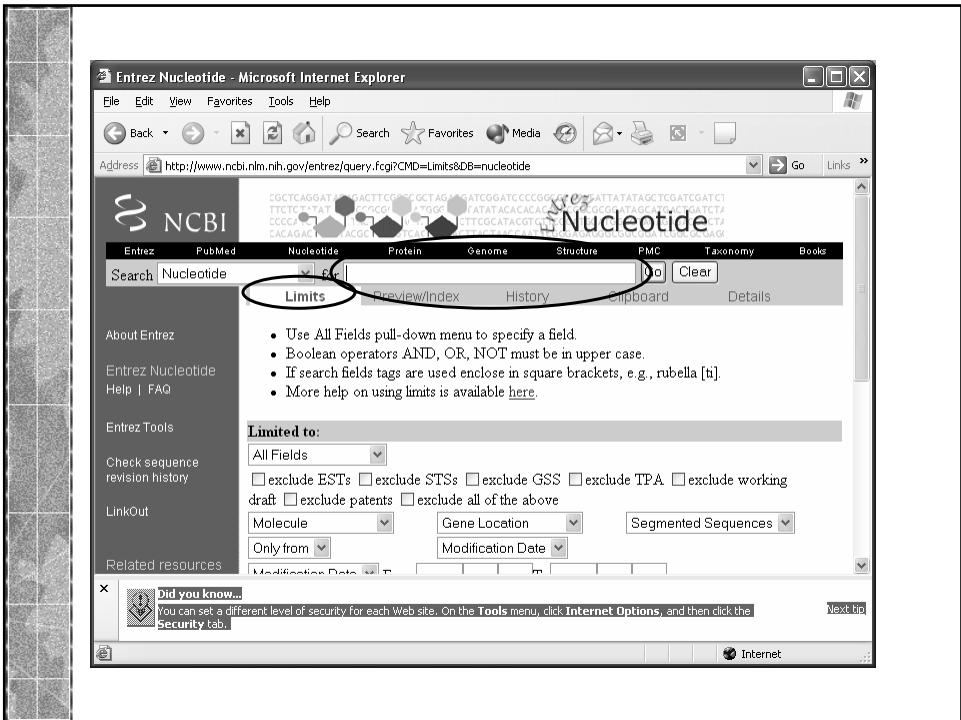
Journals: detailed information about the journals indexed in PubMed and other Entrez databases	MeSH: detailed information about NLM's controlled vocabulary
--	--

Enter terms and click 'GO' to run the search against ALL the databases, OR
 Click Database Name or icon to go directly to the Search Page for that database, OR
 Click Question Mark for a short explanation of that database.

GenBank

- ⌘ National Institute of Health, National Library of Medicine, National Center for Biotechnology Information
- ⌘ <http://www.ncbi.nlm.nih.gov/>
- ⌘ <http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html>

- ⌘ Retrieve a sequence from GenBank
- ⌘ Analyze raw sequence data
 - ◆ Base calling
 - ◆ Editing
 - ◆ Obtaining a consensus sequence
 - ◆ Translating
 - ◆ Restriction mapping
 - ◆ Similarity comparisons
 - ◆ Motif searches

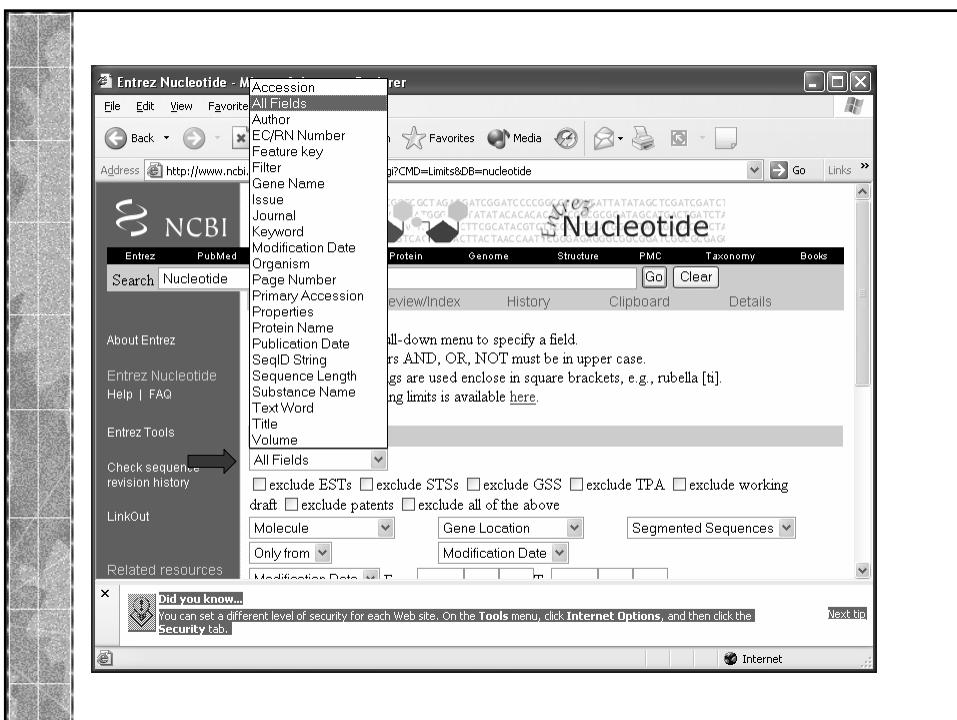
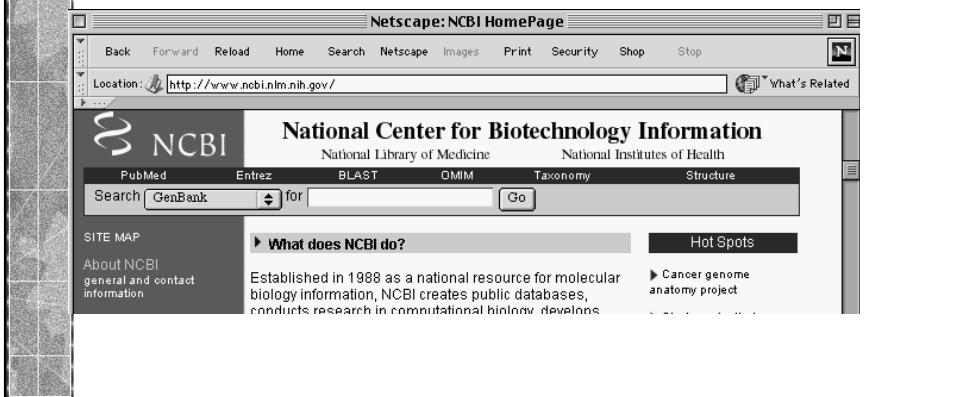


Accession Numbers!!

- ✿ Databases are designed to be searched by accession numbers (and locus IDs)
- ✿ These are guaranteed to be non-redundant, accurate, and not to change.
- ✿ Searching by gene names and keywords is inexact and retrieves more than one record usually

Type in a Query term

- Enter your search words in the query box and hit the “Go” button



Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide>

NCBI Nucleotide

Search Nucleotide for soybean AND chalcone AND synthase

Limits Preview/Index History Clipboard Details

About Entrez Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Entrez Tools

Check sequence revision history LinkOut Related resources

Limited to:

All Fields

exclude ESTs exclude STSs exclude GSS exclude TPA exclude working draft exclude patents exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide>

NCBI Nucleotide

Search Nucleotide for soybean AND chalcone AND synthase

Limits Preview/Index History Clipboard Details

About Entrez Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Entrez Tools

Check sequence revision history LinkOut Related resources

Limited to:

All Fields

exclude ESTs exclude STSs exclude GSS exclude TPA exclude working draft exclude patents exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide

NCBI Nucleotide

Search Nucleotide soybean AND chalcone synthase 6 Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

Limited to:

All Fields

exclude ESTs exclude STSs exclude GSS exclude TPA exclude working draft exclude patents exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

This screenshot shows the NCBI Entrez Nucleotide search interface in Microsoft Internet Explorer. The search bar contains the query "soybean AND chalcone synthase 6". Below the search bar are various limit options, including checkboxes for excluding ESTs, STSs, GSS, TPA, and working draft entries. There are also dropdown menus for Molecule, Gene Location, and Segmented Sequences, and a "Only from" dropdown. A "Did you know..." tip about security settings is visible at the bottom left.

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide

NCBI Nucleotide

Search Nucleotide for soybean AND chalcone synthase 6 Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

Related resources

Display: Summary Show: 20 Send to Text

Items 1-9 of 9 One page.

1: BG725328 ←
sae35d12.y1 Gm-c1051 Glycine max cDNA clone GENOME SYSTEMS CLONE ID:
Gm-c1051-7103 5' similar to SW:CHS6_SOYBN P30080 CHALCONE SYNTHASE
6;, mRNA sequence
gi|14008724|gb|BG725328.1||14008724|

2: BE346944 ←
sp33b03.y1 Gm-c1043 Glycine max cDNA clone GENOME SYSTEMS CLONE ID:
Gm-c1043-6 5' similar to SW:CHSA_IPOCO P48393 CHALCONE SYNTHASE A;,
mRNA sequence

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

This screenshot shows the results of the search "soybean AND chalcone synthase 6" on the NCBI Entrez Nucleotide site. Two results are displayed: one for "BG725328" and another for "BE346944". Each result includes a checkbox, the clone ID, the organism (Glycine max), the clone name (e.g., "GENOME SYSTEMS CLONE ID: Gm-c1051-7103"), a similarity statement, and a link to the full sequence record (gi|14008724|gb|BG725328.1||14008724| and gi|14008724|gb|BE346944.1||14008724|). A "Did you know..." tip about security settings is visible at the bottom left.

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Go Links

Address: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for | Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File

I: BG725328 sae35d12.y1 Gm-c1. [gi|14008724]

IDENTIFIERS

dbEST Id: 8490125
EST name: sae35d12.y1
GenBank Acc: BG725328
Genbank gi: 14008724

CLONE INFO

Clone Id: GENOME SYSTEMS CLONE ID: Gm-c1051-7103 (5')
DNA type: cDNA

PRIMERS

PolyA Tail: Unknown

SEQUENCE

GGTTGTGCCAAGTCCATTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAAACAGCAC
ACATGGACTTGACATGTTACCATACTGGCTAACGACGGTCTAGTGGCTTCATTTTTC
ATGCTTCATCTCAACTTAGCCTAACCTGGTCAAATTGCTGGTCCACCAGGGTGTGC
AATCCAAAAGATAAGAGTTGTAATCATCAATTTCATAAGGTTGAAGGCTTCACCAAAAGGC
CTTTTGGATGTTCTTGAGAGTAGCTTGGCAGGAACATCTTGGAGAGATGAAAGTGTCTC
TACTTTGGCGAAGGTGGCATCAATTGGCTTCAGCTGGAGGGATGTTTGTGAGT
CCACACAAAGCTCAACAAAAGGCTTTCAAGCTGGAGAGATCTGATCCAACAAATGACAGC
AGCTGCAACCATCTCAACAAAGGCTTGCCCCACAAGGCTGTCAAGATGTGTCACTCGG
GCCACGAAATGACTGCTGTGATCTCGA

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Go Links

Address: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724

SEQUENCE

GGTTGTGCCAAGTCCATTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAAACAGCAC
ACATGGACTTGACATGTTACCATACTGGCTAACGACGGTCTAGTGGCTTCATTTTTC
ATGCTTCATCTCAACTTAGCCTAACCTGGTCAAATTGCTGGTCCACCAGGGTGTGC
AATCCAAAAGATAAGAGTTGTAATCATCAATTTCATAAGGTTGAAGGCTTCACCAAAAGGC
CTTTTGGATGTTCTTGAGAGTAGCTTGGCAGGAACATCTTGGAGAGATGAAAGTGTCTC
TACTTTGGCGAAGGTGGCATCAATTGGCTTCAGCTGGAGGGATGTTTGTGAGT
CCACACAAAGCTCAACAAAAGGCTTTCAAGCTGGAGAGATCTGATCCAACAAATGACAGC
AGCTGCAACCATCTCAACAAAGGCTTGCCCCACAAGGCTGTCAAGATGTGTCACTCGG
GCCACGAAATGACTGCTGTGATCTCGA

Quality: High quality sequence stops at base: 400

Entry Created: May 8 2001
Last Updated: Jul 22 2004

COMMENTS

When it has been determined, an EST from the other end of this clone is listed in the 'Other ESTs on clone' field. Possible reversed clone: similarity on wrong strand This clone is available through: Biogenetic Services, 801 32nd Ave. Brookings, SD 57006 USA (phone: 800 423 4163; email: info@biogeneticservices.com)

PUTATIVE ID

Assigned by submitter
SW:CH56_SOYB_P30080 CHALCONE SYNTHASE 6 ;

LIBRARY

Lib Name: Gm-c1051
Organism: Glycine max
Cultivar: Corolla
Tissue type: floral meristematic mRNA
Is host: no

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print E-mail Links

Address: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724

LIBRARY

Lib Name: Gm-c1051
Organism: Glycine max
Cultivar: Corolla
Tissue type: floral meristematic mRNA
Lab host: DH10B
Vector: pBluescript II SK+
R. Site 1: EcoRI
R. Site 2: XbaI
Description: The cDNA library was constructed from floral meristematic mRNA provided by Dr. Halina Knap of Clemson University. Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a XbaI restriction site. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by XbaI digestion. The cDNA fragments were directionally cloned into the EcoRI-XbaI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into DH10B host cells (GibcoBRL). This library was constructed in the laboratory of Dr. Randy Shoemaker.

SUBMITTER

Name: Shoemaker R/Public Soybean EST Project
Lab: Public Soybean EST Project
Institution: Washington University School of Medicine
Address: 4444 Forest Park Parkway, Box 8501, St. Louis, MO 63108, USA
Tel: 314 286 1800
Fax: 314 286 1810
E-mail: est@watson.wustl.edu

CITATIONS

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

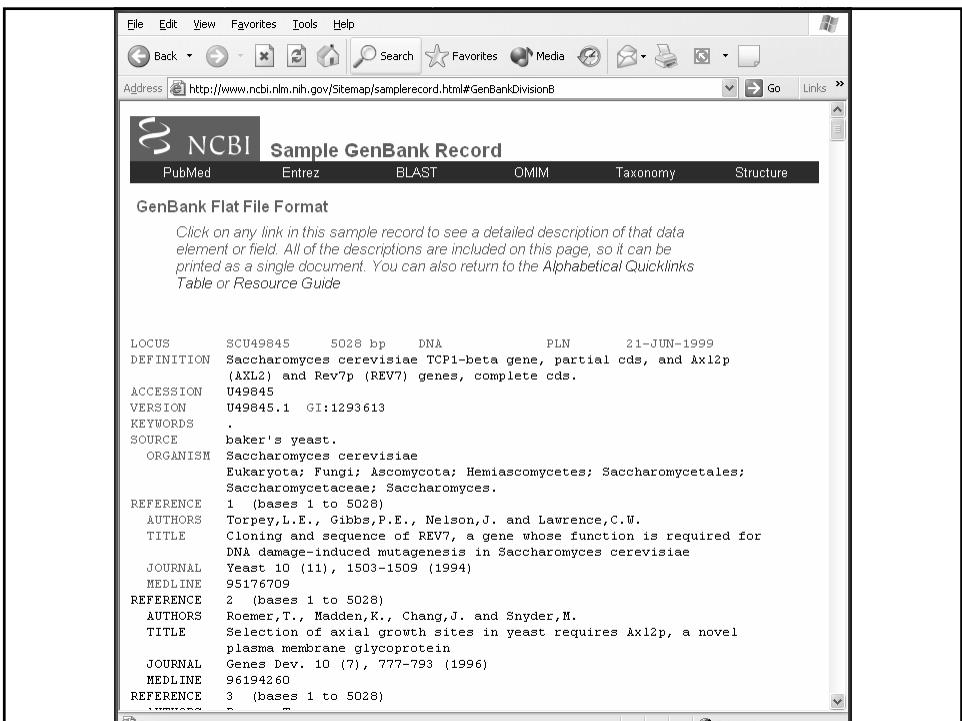
Next tip Internet

GenBank Records

- ★ Databases are composed of records
- ★ Flat File Format
- ★ Provides information
- ★ Standard, consistent organization of data

Flat file format

- ⌘ Organized in a structured manner
- ⌘ One big file
- ⌘ Large body of information assembled and distributed in consistent format
- ⌘ Lack support for procession transactions (inserts and updates)



The screenshot shows a Microsoft Internet Explorer window displaying a sample GenBank record. The title bar reads "NCBI Sample GenBank Record". The main content area is titled "GenBank Flat File Format" and contains the following text:

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide.

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS .
SOURCE baker's yeast.
ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (bases 1 to 5028)
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for
DNA damage-induced mutagenesis in *Saccharomyces cerevisiae*
JOURNAL Yeast 10 (11), 1503-1509 (1994)
MEDLINE 95176709
REFERENCE 2 (bases 1 to 5028)
AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE Selection of axial growth sites in yeast requires Axl2p, a novel
plasma membrane glycoprotein
JOURNAL Genes Dev. 10 (7), 777-793 (1996)
MEDLINE 96194260
REFERENCE 3 (bases 1 to 5028)

Some Fields of GenBank Record

- | | |
|----------------------------|--------------------------|
| ⌘ Locus Name | ⌘ Reference |
| ⌘ Sequence length | ⌘ Authors |
| ⌘ Molecule type | ⌘ Title |
| ⌘ Definition | ⌘ Journal |
| ⌘ GenBank accession number | ⌘ Medline |
| ⌘ Version | ⌘ Other references |
| ⌘ Keywords | ⌘ features |
| ⌘ Source | ⌘ Amino acid translation |
| ⌘ Organism | ⌘ Nucleotide sequence |
| ⌘ Reference | |

NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

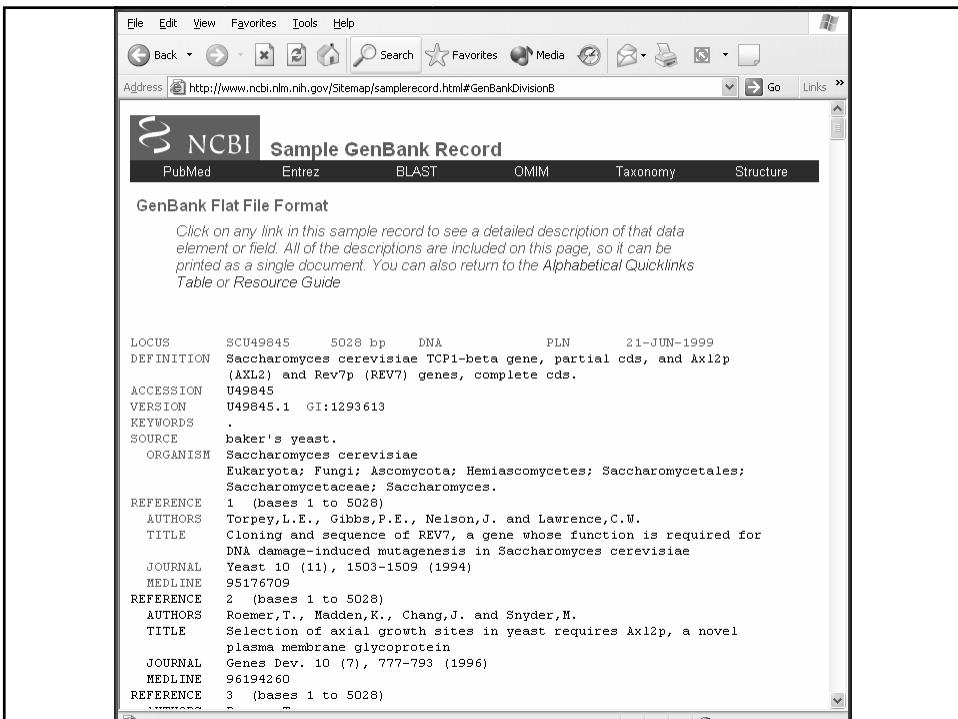
LOCUS	SCU49845	5028 bp	DNA	PLN	21-JUN-1999
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Ax12p (AXL2) and Rev7p (REV7) genes, complete cds.				
ACCESSION	U49845				
VERSION	U49845.1 GI:1293613				
KEYWORDS	.				
SOURCE	baker's yeast.				
ORGANISM	Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.				
REFERENCE	1 (bases 1 to 5028)				
AUTHORS	Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.				
TITLE	Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in <i>Saccharomyces cerevisiae</i>				
JOURNAL	Yeast 10 (11), 1503-1509 (1994)				
MEDLINE	95176709				
REFERENCE	2 (bases 1 to 5028)				
AUTHORS	Roemer,T., Madden,K., Chang,J. and Snyder,M.				
TITLE	Selection of axial growth sites in yeast requires Ax12p, a novel plasma membrane glycoprotein				
JOURNAL	Genes Dev. 10 (7), 777-793 (1996)				
MEDLINE	96194260				
REFERENCE	3 (bases 1 to 5028)				

Locus Name

Unique

- ⌘ Up to 10 characters
- ⌘ 6 character
 - ◆ Genus species
- ⌘ 8 characters
 - ◆ Just accession number

- ⌘ Better to search for accession number than Locus Name



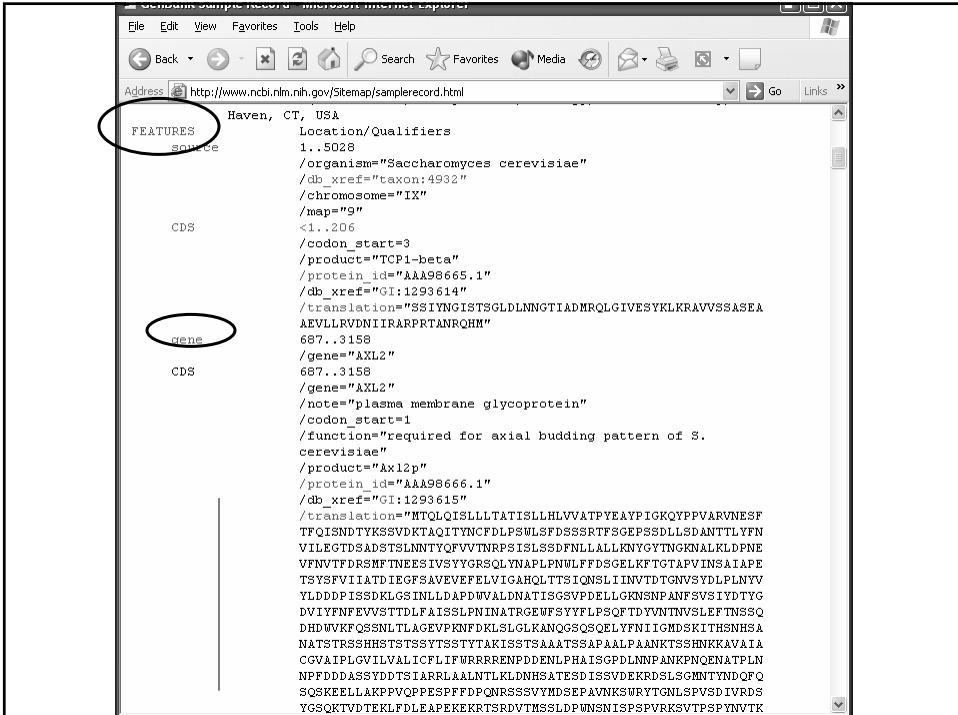
The screenshot shows a web browser window displaying a sample GenBank record from NCBI. The URL in the address bar is <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#GenBankDivisionB>. The page title is "NCBI Sample GenBank Record". The main content area is titled "GenBank Flat File Format" and contains the following text:

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the [Alphabetical Quicklinks Table](#) or [Resource Guide](#).

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS .
SOURCE baker's yeast.
ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (bases 1 to 5028)
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for
DNA damage-induced mutagenesis in *Saccharomyces cerevisiae*
JOURNAL Yeast 10 (11), 1503-1509 (1994)
MEDLINE 95176709
REFERENCE 2 (bases 1 to 5028)
AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE Selection of axial growth sites in yeast requires Axl2p, a novel
plasma membrane glycoprotein
JOURNAL Genes Dev. 10 (7), 777-793 (1996)
MEDLINE 96194260
REFERENCE 3 (bases 1 to 5028)

GenBank Accession Number

- ⌘ Unique identifier for sequence record
- ⌘ Usually a combination of letter(s) and numbers
- ⌘ Do not change even if information changes
- ⌘ Newer accession numbers to new submission using some of this data



The screenshot shows a Microsoft Internet Explorer window displaying a GenBank XML record. The URL in the address bar is <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>. The XML code is as follows:

```
<?xml version="1.0"?>
<!DOCTYPE Sequences SYSTEM "http://www.ncbi.nlm.nih.gov/Sequences.dtd">
<Sequences>
  <Source db_xref="taxon:4532" organism="Saccharomyces cerevisiae" map="" name="Saccharomyces cerevisiae" taxid="4532" taxname="Saccharomyces cerevisiae">
    <Location Qualifiers="source">Haven, CT, USA</Location>
    <Features>
      <Feature db_xref="GI:1293614" feature="CDS" location="1..5028" product="TCP1-beta" protein_id="AAA98665.1" start="1" type="CDS">
        <Location Qualifiers="source">Haven, CT, USA</Location>
        <Translation>SSINYGISTSGLDLNNGTIADMRLQLGIVESYKLKRAVVSSASEA</Translation>
        <Translation>AEVLLRVNDNIIRARPTANRQHM</Translation>
        <Location Qualifiers="source">Haven, CT, USA</Location>
        <Translation>687..3158</Translation>
        <Gene db_xref="GI:1293615" name="AXL2" start="3158" type="gene">
          <Location Qualifiers="source">Haven, CT, USA</Location>
          <Note>plasma membrane glycoprotein</Note>
          <Function>required for axial budding pattern of S. cerevisiae</Function>
          <Product>Ax12p</Product>
          <Protein db_xref="GI:1293615" id="AAA98666.1" name="AXL2" start="3158" type="protein">
            <Location Qualifiers="source">Haven, CT, USA</Location>
            <Translation>MTQLQISLLTATISLLHLVVATPYEAYPIKGQYPVARVNESF</Translation>
            <Translation>TFQISNDTYKSSVUDKTA01TYNCFDLPSWLSFSDSSRTFSCEPSSDLSANTLYPN</Translation>
            <Translation>VILEGTDSDADTSLLNNTTQFWVTINPPSISLSSSDFNLLALLKNGVYTNGENALKLDPNE</Translation>
            <Translation>VENVTFDRSMFTTHEESIVSYGRSQLNAPLNULFFDSGELKFTGTAPVINSIAPE</Translation>
            <Translation>TSYSFVIIATDIEGFSAVEVEFELVIGAHQHLTTSI0NSLIINVTDTGWSVYDPLPLNVY</Translation>
            <Translation>YLDPP135DKLGSINLLDADPVALDNATISGVDPDELLCKNSNPANF5VS1YDITYG</Translation>
            <Translation>DVIFYNFEEVSTTDLFAISSLNPINATRGWEMFSYVLPFQFTDVNTMVSLEFTMSSQ</Translation>
            <Translation>DHDWVKFQSNTLTLAGEVPKNFDKLSLGLKANQGSOSQELYFNIIGIDMSKITHSNHSANAT</Translation>
            <Translation>NATSTRSSHSTSTSSYTSYTAKISSTSAAATSSAAPALPAANKTSSHNNKKAVATA</Translation>
            <Translation>CGVAIPLGVLVALICFLIFURRRRENPDDENLPHAIISGPDLNWPANKPQNEMATPLN</Translation>
            <Translation>NPFFDDASSYYDTSIARRLAALNTLKDLDNHSATESDISSVDEKRDSLGHMNTYNDQFQ</Translation>
            <Translation>SQSEKEELLAKRPVQOPPFDPQNRSSSVYMDSEPAVNEKSWRYTGNLSPVSDIVRDS</Translation>
            <Translation>YGSQKTVDTKELFDELAPEKEKRTSRDVTHMSSLDPUWNSMISPPSPVRKSVPSPYNTVK</Translation>
          </Protein>
        </Gene>
      </Feature>
    </Features>
  </Source>
</Sequences>
```

Address : http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

```

NATSTRSSHSTSTSSYTAKISSTSAAATSSAPAALPAANKTSSHNNKAVAI
CGVAIPLGVILVALICFLIFWRRRENPDDENLPHAIISGPDLNPNPKQNENATPLN
NPFDDASSYDTSIARRLAALNTLKLDNHSATESDISSVDERKDSLSCMNNTYNDQFQ
SQSKEELLAKPVOPPESPFPDPQRSSSVYMDSEPAVNKSURYTGNLSPVSDIVRDS
YGSQKTVDTEKLFDEAPEKEKRTSRDVTNSSLDPWNNSNISPSPVRKSVPSPYNTVK
HRNRHLQNIQDQSQSGKNGITPTTMSTSSDDFVPKDGENFCWVHSMEPDRRPSKKRL
VDFSNKSNVNVGQVKDIHGRIPLEM"
complement (3400..4037)
/gene="REV7"
complement (3400..4037)
/gene="REV7"
/codon_start=1
/product="Rev7p"
/protein_id="AAA98667.1"
/db_xref="GI:1293616"
/translation="MNRWVEKWLRLVYLKCYINLILFYRNVPPQSFDYTTTYSFNLQP
FVPIRHHPALIDYIEELILDVLSLRKTVYRFSCICINKNDLIEKWLDFSELQHVD
KDDQIITTEVFDDEFRSSLNSLIMHLEKLPVNDDTITFEAVINAELELGKLDRN
RVDLSLEEKAEIERDSNWVKCQEDENLDPNNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
LISGDDKILNGVYSQEEGESIFGSF"
BASE COUNT      1510  a    1074  c    835  g    1609  t
ORIGIN
1 gatcctccat atacaacggat atctccacat cagggtttaga ttcacaaac ggaaccattg
61 ccgcacatgg acagtttaggt atcgctgaga gttacaagct aaaacggacca gtatgcacgt
121 ctgcacatgtg aagccgtgaa gttctactaa gggtggataaa catcatccgt gcaagaccaa
181 gaaccgccaa tagacacat atgtacataat tttaggatat acctcgaaaaaa taataacccg
241 ccacactgtc atttataataa tttagaaacag aacgcacaaaat ttatccacta tataattccaa
301 agacgcgaaa aaaaaaaggac aacgcgtcat agaacttttg gcaatttcgcg tcacaatataa
361 attttgcacaa cttatgttgc cttttcgacg agtacttcgag cccttgcctca agaatgttaat
421 aataccaccat gtaggttatgg ttaaaagatag catctccaca acctccaaagc ttcttgcgcga
481 gagtcgcgcctt cttttgtcga gtaattttca cttttcatat gagaactttat ttttttatc
541 tttaactctca catcctgtcg tgatggacac tgcaacacgc accatccatc gaaagacacgaa
601 aacaatattctt aataaaaaaa ttatattcttc ctgcacacgc ttccctgttcc caacatctta
661 cgatattccaa gaagcatccat cttaccatcgatc cttaccatcgatc gatttccatc ttgtgtgacag
721 ctactatatac actactccat ctatgtatgg ccacgcctta tgaggcatat cttatccggaa
781 aacaatacc cccatgtggca agagtcaatg aatcggtttac atttcaatatttccatgtata
841 cttataatccat gttgttagac aagacagtc aataaaatcatc caatgttcgac gacttacccgaa
901 gctggcttc gtttgactt agttcttagaa cgtttctcagg tgaaccttc tctgtacttac

```

Refine the Query

- ⌘ Often a search finds too many (or too few) sequences, so you can go back and try again with more (or fewer) keywords in your query
- ⌘ The “History” feature allows you to combine any of your past queries.
- ⌘ The “Limits” feature allows you to limit a query to specific organisms, sequences submitted during a specific period of time, etc.
- ⌘ [Many other features are designed to search for literature in MEDLINE]

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nucleotide>

NCBI Nucleotide

Search Nucleotide for Limits

About Entrez Entrez Nucleotide Help | FAQ Entrez Tools Check sequence revision history LinkOut Related resources

Limits

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available [here](#).

Limited to:

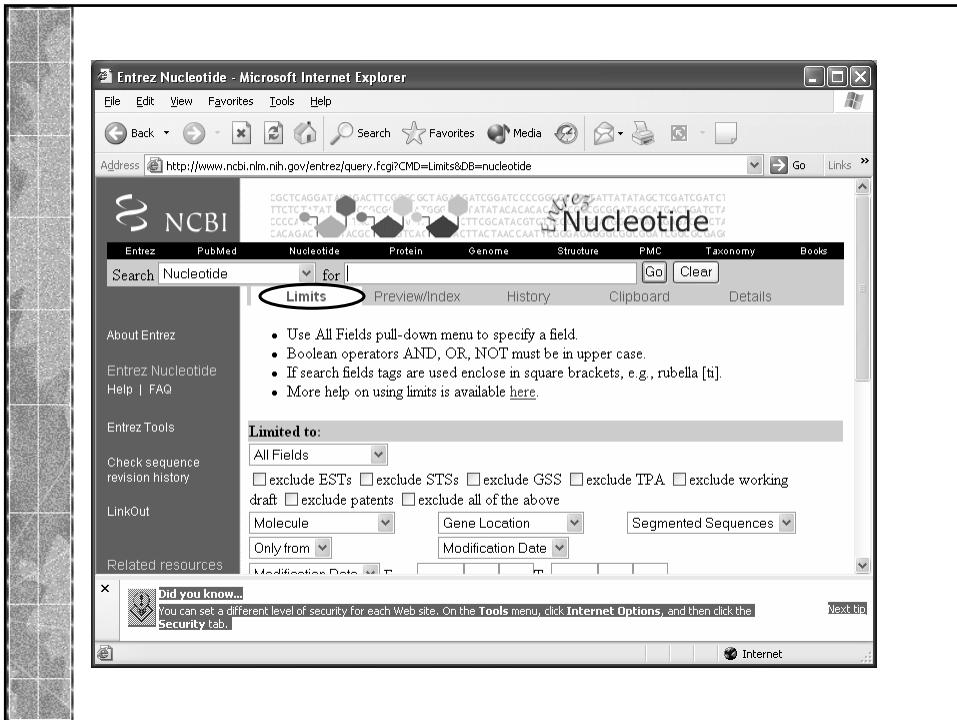
All Fields exclude ESTs exclude STSs exclude GSS exclude TPA exclude working draft exclude patents exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet



Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Index&DB=nucleotide>

NCBI Nucleotide

Search Nucleotide for Preview

About Entrez Entrez Nucleotide Help | FAQ Entrez Tools Check sequence revision history LinkOut Related resources

Preview/Index

- Enter terms and click Preview to see only the number of search results.
- To combine searches use # before search number, e.g., (#2 OR #3) AND asthma. No history available

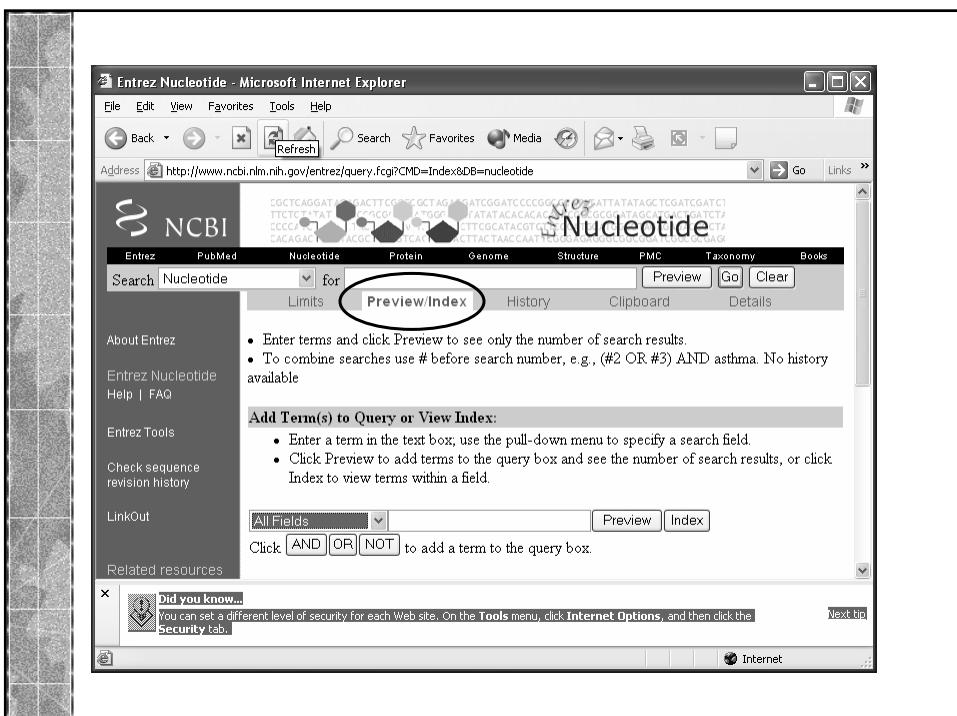
Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.

All Fields
Click **AND** **OR** **NOT** to add a term to the query box.

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet



Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Index&DB=nucleotide

Search Favorites Media Links

Back Forward Stop Refresh Home Search

Accession All Fields Author EC/RN Number Feature key Filter Gene Name

NCBI Nucleotide

Entrez PubMed Search Nucleotide About Entrez Entrez Nucleotide Help | FAQ Entrez Tools Check sequence revision history LinkOut

Protein Genome Structure PMC Taxonomy Books

view/Index History Clipboard Details

Preview Go Clear

Preview to see only the number of search results.
use # before search number, e.g., (#2 OR #3) AND asthma. No history

or View Index:
the text box; use the pull-down menu to specify a search field.
add terms to the query box and see the number of search results, or click
within a field.

All Fields Preview Index

Click AND OR NOT to add a term to the query box.

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

Related resources

Entrez Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=History&DB=nucleotide

Search Favorites Media Links

Back Forward Stop Refresh Home Search

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for soybean AND chalcone synthase AND full-length Preview Go Clear

Limits Preview/Index History Clipboard Details

about Entrez Entrez Nucleotide Help | FAQ Entrez Tools Check sequence revision history LinkOut

Related resources LAST Reference sequence project Search for Genes Submit to GenBank Search for full length DNAs

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

Search Most Recent Queries Time Result

#2 Search soybean AND chalcone synthase AND full-length 10:16:02 30

#1 Search soybean AND chalcone synthase 10:14:19 339

Clear History

LAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length DNAs

DNA similarity search

Find related sequences

- * Find ESTs
 - ◆ If not full-length, may allow assembly from ESTs
- * Find other family members
 - ◆ Organization and function
- * Find similar genes from other cultivars
 - ◆ SNP discovery
- * Find similar genes from other organisms
 - ◆ Phylogenetic relationships

ESTs (Expressed Sequence Tags)

- * partial cDNA sequences
- * dbEST at NCBI
 - a comprehensive set of all public EST data
- * UniGene at NCBI
 - clusters of ESTs and known genes from key species
 - does NOT have consensus sequences
 - has far too many clusters to be representative of real genes (129 K human clusters)

Find related DNA sequences

- * Similarity Search (BLAST)
- * NCBI GenBank database

BLAST Searches

- ⌘ Compare your sequence with database
- ⌘ <http://www.ncbi.nlm.nih.gov/BLAST/>
- ⌘ Nucleotide
- ⌘ Protein
- ⌘ Targeted to a genome

⌘ BLAST

- Basic Local Alignment Search Tool
- Local alignment
- Tutorial at:
<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

BLAST

★ Discontiguous Mega BLAST

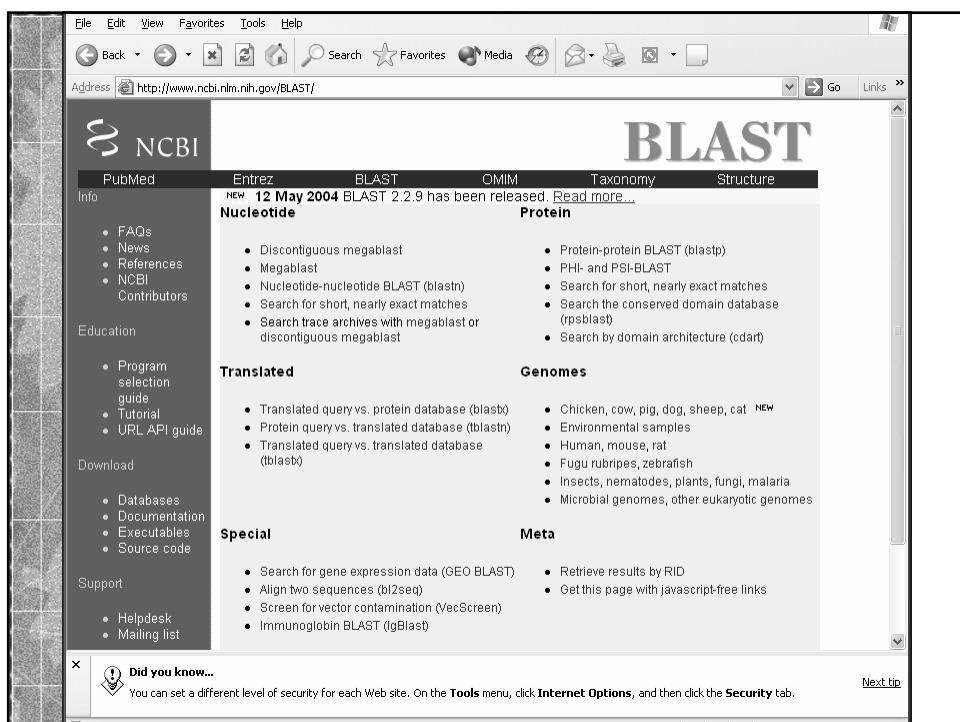
- Comparison of diverged sequences especially from different organisms
- Alignments with low degree of identity
- Looks for hits in “non-consecutive positions”

★ Mega BLAST

- Slight differences in similarity
- Not effective at low degree of identity
- Faster; handles longer sequences

★ BLAST

- Local alignment tool
- <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>



The screenshot shows the NCBI BLAST homepage. The top navigation bar includes links for File, Edit, View, Favorites, Tools, and Help. The address bar shows the URL <http://www.ncbi.nlm.nih.gov/BLAST/>. The main header features the NCBI logo and the word "BLAST". A banner at the top right states "NEW 12 May 2004 BLAST 2.2.9 has been released. [Read more...](#)". The page is divided into several sections:

- PubMed**: Links to FAQs, News, References, NCBI Contributors.
- Info**: Links to Program selection guide, Tutorial, URL API guide.
- Education**: Links to Databases, Documentation, Executables, Source code.
- Download**: Links to Helpdesk, Mailing list.
- Special**: Links to GEO BLAST, bl2seq, VecScreen, IgBlast.
- Translated**: Links to Translated query vs. protein database (blast), Protein query vs. translated database (tblastn), Translated query vs. translated database (tblastx).
- Nucleotide**: Links to Discontiguous megablast, Megablast, Nucleotide-nucleotide BLAST (blastn), Search for short, nearly exact matches, Search trace archives with megablast or discontiguous megablast.
- Protein**: Links to Protein-protein BLAST (blastp), PHLI- and PSI-BLAST, Search for short, nearly exact matches, Search the conserved domain database (rpsblast), Search by domain architecture (cdart).
- Genomes**: Links to Chicken, cow, pig, dog, sheep, cat, Environmental samples, Human, mouse, rat, Fugu rubripes, zebrafish, Insects, nematodes, plants, fungi, malaria, Microbial genomes, other eukaryotic genomes.
- Taxonomy**: Links to Retrieve results by RID, Get this page with javascript-free links.
- Structure**: Links to various structural biology resources.

A "Did you know..." sidebar at the bottom left provides information on setting security levels for web sites. A "Next tip" link is located at the bottom right.

This screenshot shows a Microsoft Internet Explorer window displaying the NCBI BLAST Nucleotide search interface. The address bar shows the URL <http://www.ncbi.nlm.nih.gov/BLAST/>. The page header includes the NCBI logo and the word "Nucleotide". On the left, there's a sidebar with links for FAQs, News, References, NCBI Contributors, and Education. The main content area has a heading "Did you know..." with a tip about security settings. Below it, under the "Nucleotide" section, is a list of search options. Two specific items in this list are highlighted with black arrows pointing to them:

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

This screenshot shows a Microsoft Internet Explorer window displaying the main NCBI BLAST homepage. The address bar shows the URL <http://www.ncbi.nlm.nih.gov/BLAST/>. The page features a large "BLAST" logo at the top. A banner at the top right indicates "NEW 12 May 2004 BLAST 2.2.9 has been released. [Read more...](#)". The left sidebar contains links for PubMed, Entrez, OMIM, Taxonomy, Structure, and various sections like Info, Education, Download, Support, and Special. The main content area is divided into several sections: "Discontiguous megablast" (which is circled in red), "Translated", "Genomes", "Special", and "Meta". Each section lists specific search types or databases. At the bottom of the page, there's a "Did you know..." box with a tip about security settings.

You have a sequence.
Does it have similarity to other known genes???
Copy DNA sequence from file

```
GGTTGTCCAAGTCCATTCTATTGACTTCTCCTCATTTGATCCAAGATGAACAGCAC  
ACATGCACCTGACATGTTACCATACTCGCTAACGCACGTGCTAGTAGCTTCATTTC  
ATGCTTCATCCTAACTTAGCCTCAACTGGTCCAAAATTGCTGGTCCACCAGGGTGTGC  
AATCCAAAAGATAGAGTTGAATCATCAATTCTAAGGGTTGAAGGCTTCAACCAAGGC  
CTTTTCGATGTTCTTGAGATGAGTCCAGGAACATCCTTGAGGAGATGAAAGTGACTCC  
TACTTGGCGAAGGTGGCCATCAATAGCGCCTTCGCTGTGAAAGGATTGTTGTGCAGT  
CCACACAAGCTCAAACAAAGGCTTCAGCTGGCAGAGGATCTGATCCAACAATGACAGC  
GCTGCACCATCTCCAAACAAGGCTTCCCCACAAGGCTGTCAAGATGTGTCACTCGG  
GCCACGAAATGTGACTGCTGTGATCTCCGA
```

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&c

megablast BLAST

Nucleotide Protein Translations Retrieve results for an RID

What is discontiguous Mega BLAST?

Search

Load query file from disk Browse...

Set subsequence From: To:

Choose database nr

Return alignment endpoints only

Now: **BLAST!** or **Reset query** **Reset all**

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Links

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&c

NCBI megablast BLAST

Nucleotide Protein Translations Retrieve results for an RID

What is discontiguous Mega BLAST?

Search: GGTTGTGCCAAGTCCATTTCATTTGACTTCTTCCTCATTTGATCCAAGATGAACAGCA
CACATGCACTTGACATGTTACCATCTCGCTAACGCACGTGCTAGTAGCTTCATTTT
TCATGCTTCAACTCTAACTTAGCCTCAACTGGTCCAAAATTGCTGGTCCACCAGGGTG
TGCATTCAAAAGATAGAGTTGTAATCATCAATTCTAAAGGGTTGAAGGCTTCACCA
AGGCCTTTGATGTTCTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAGTG

Load query file from disk: [Browse...]

Set subsequence From: [] To: []

Choose database: nr [▼]

Return alignment endpoints only: []

Now: **BLAST!** or **Reset query** **Reset all**

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Links

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&c

NCBI megablast BLAST

Nucleotide Protein Translations Retrieve results for an RID

What is discontiguous Mega BLAST?

Search: GGTTGTGCCAAGTCCATTTCATTTGACTTCTTCCTCATTTGATCCAAGATGAACAGCA
CACATGCACTTGACATGTTACCATCTCGCTAACGCACGTGCTAGTAGCTTCATTTT
TCATGCTTCAACTCTAACTTAGCCTCAACTGGTCCAAAATTGCTGGTCCACCAGGGTG
TGCATTCAAAAGATAGAGTTGTAATCATCAATTCTAAAGGGTTGAAGGCTTCACCA
AGGCCTTTGATGTTCTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAGTG

Load query file from disk: [Browse...]

Set subsequence From: [] To: []

Choose database: nr [▼]

nr
est
est_human
est_mouse
est_others
gss
htgs
pat
pdb
month
slu_repeats
dbsts
chromosome
wgs
env_nt

Return alignment endpoints only: []

Now: **Reset query** **Reset all**

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto&from_disk [BROWSE... History]

Set subsequence From: _____ To: _____

Choose database est_others ↗

Return alignment endpoints only

Now: **BLAST!** or **Reset query** **Reset all**

Options for advanced blasting

Limit by entrez query **Glycine max** ↗ or select from All organisms

Choose filter Low complexity Human repeats Mask for lookup table only Mask lower case

Expect **10**

Word Size **11**

Percent Identity, match_mismatch **None, 1. -2** scores

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

formatting BLAST

Nucleotide Protein Translations Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = (509 letters)

Your search was limited by an Entrez query: Glycine max

The request ID is **1090952418-29109-8211948582.BLASTQ4**

Format ↗ or **Reset all**

The results are estimated to be ready in 23 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show Graphical Overview Linkout Sequence Retrieval NCBI-21 Alignment **HTML** format

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

Internet

NCBI *formatting BLAST*

Nucleotide	Protein	Translations	Retrieve results for an RID
----------------------------	-------------------------	------------------------------	---

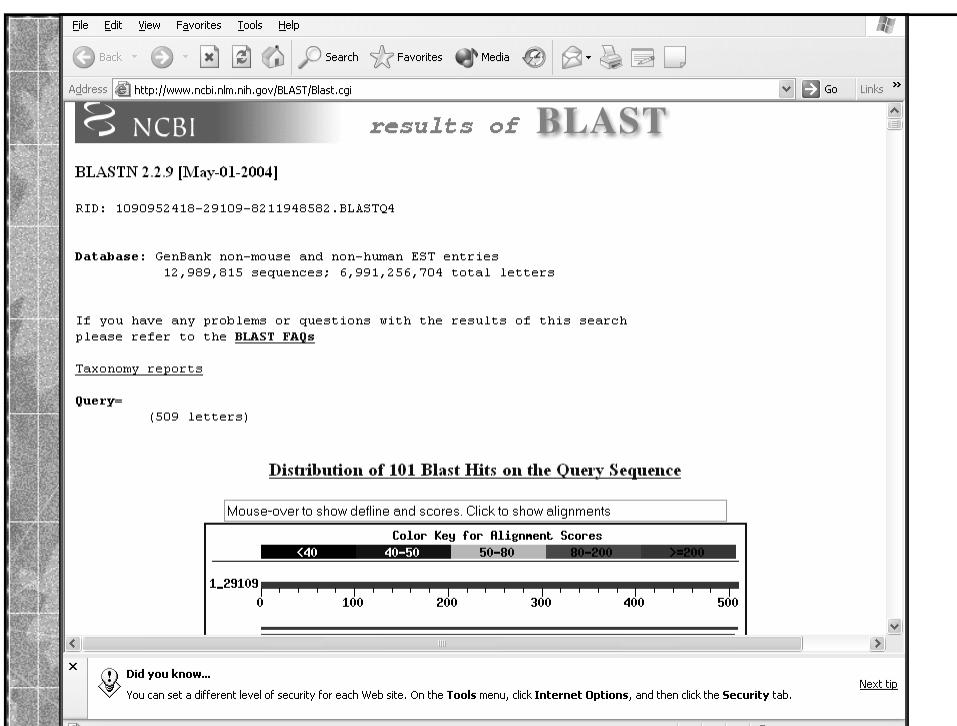
Your request has been successfully submitted and put into the Blast Queue.
Query = (1509 letters)
Your search was limited by an Entrez query: Glycine max

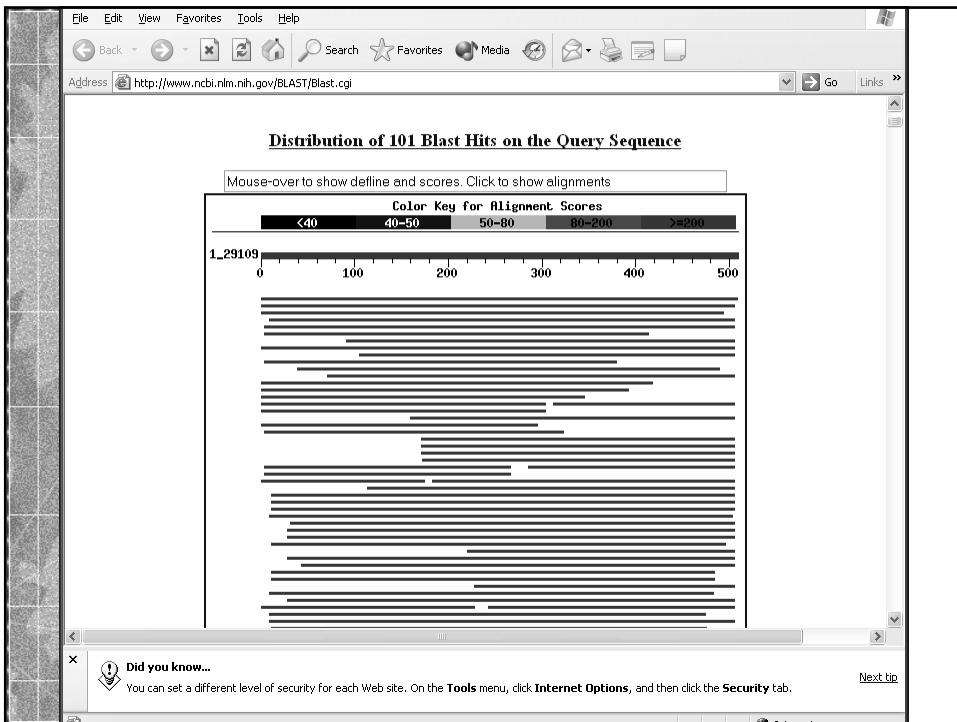


The request ID is

Format! or **Reset all**

The results are estimated to be ready in 37 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

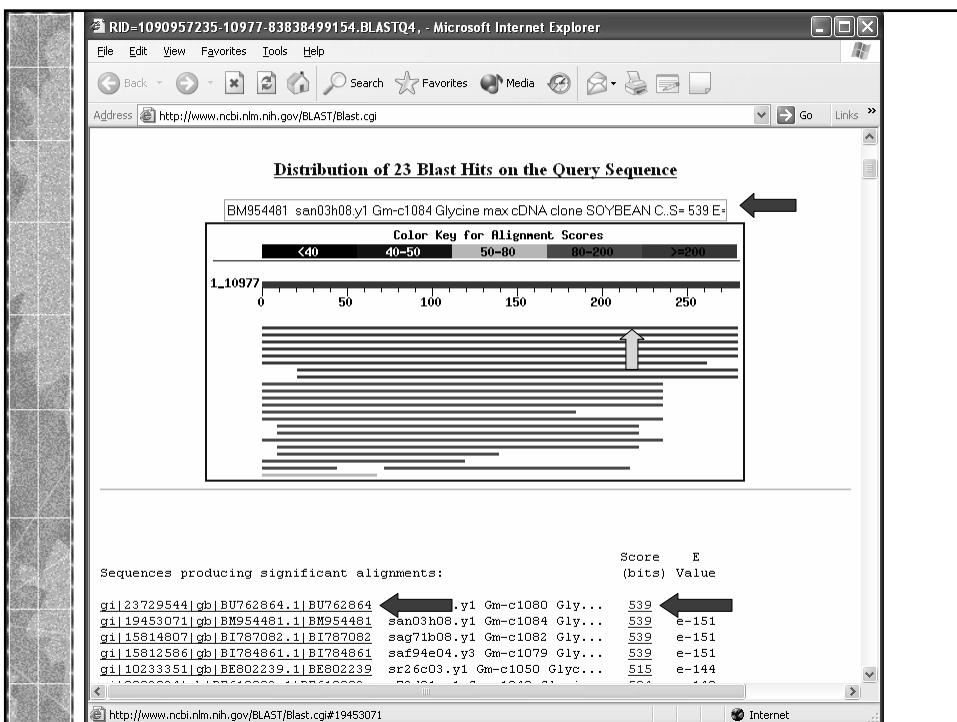
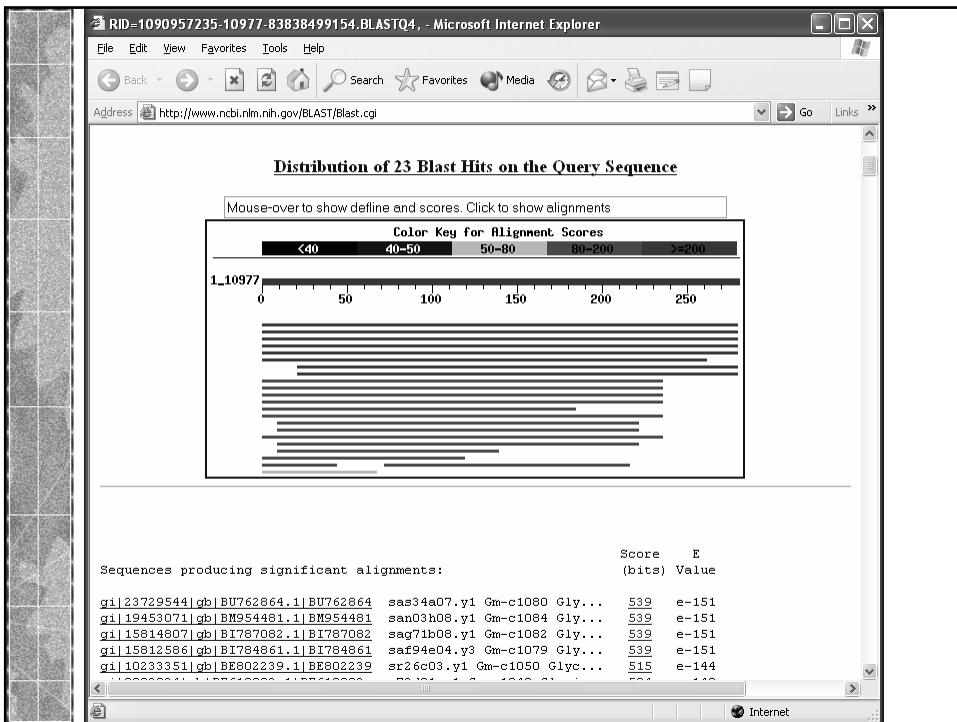




Sequences producing significant alignments:

	(bits)	Value
gi 14008724 gb BG725328.1 BG725328	sae35d12.y1 Gm-c1051 Gly...	965 0.0 U
gi 127427571 gb CA939091.1 CA939091	sav41g09.y1 Gm-c1069 Gly...	712 0.0 U
gi 16346573 gb BI972168.1 BI972168	sag88a12.y1 Gm-c1084 Gly...	694 0.0 U
gi 152874781 gb BI471369.1 BI471369	sag19f06.y1 Gm-c1080 Gly...	694 0.0 U
gi 137995974 gb CF807563.1 CF807563	psHB025x009f USDA-IFAFS:...	687 0.0 U
gi 1237322961 gb BU764307.1 BU764307	sar98d05.y2 Gm-c1080 Gly...	675 0.0 U
gi 137994162 gb CF805908.1 CF805908	psHB001x006f USDA-IFAFS:...	664 0.0 U
gi 16342613 gb BI968208.1 BI968208	GM830004B12FO5 Gm-r1084 ...	658 0.0 U
gi 199347251 gb B0079755.1 B0079755	san17h09.y1 Gm-c1084 Gly...	642 0.0 U
gi 19934445 gb BQ079475.1 BQ079475	san14c01.y1 Gm-c1084 Gly...	625 e-176 U
gi 13478388 gb BG507884.1 BG507884	sac82e09.y1 Gm-c1072 Gly...	623 e-176 U
gi 137996764 gb CF808353.1 CF808353	psHB034xJ14f USDA-IFAFS:...	592 e-166 U
gi 16106094 gb BI893834.1 BI893834	sag93f11.y1 Gm-c1084 Gly...	587 e-165 U
gi 19935253 gb BQ080283.1 BQ080283	san31a09.y1 Gm-c1084 Gly...	565 e-158 U
gi 11412994 gb BF425005.1 BF425005	su53b09.y1 Gm-c1069 Glyc...	496 e-138 U
gi 379966661 gb CF808255.1 CF808255	psHB033xI01f USDA-IFAFS:...	465 e-128 U
gi 19934418 gb BQ079448.1 BQ079448	san13h02.y1 Gm-c1084 Gly...	465 e-128 U
gi 137996212 gb CF807801.1 CF807801	psHB028xG08f USDA-IFAFS:...	450 e-124 U
gi 13562975 gb BG551195.1 BG551195	sad34d04.y1 Gm-c1074 Gly...	448 e-123 U
gi 17673732 gb AM160139.1 AM160139	pblt17 soybean, century c...	442 e-121 U
gi 15815451 gb BI787726.1 BI787726	sag75a07.y1 Gm-c1084 Gly...	435 e-119 U
gi 13480079 gb BG509422.1 BG509422	sad13f03.y1 Gm-c1074 Gly...	435 e-119 U
gi 137996627 gb CF808216.1 CF808216	psHB033xC14f USDA-IFAFS:...	429 e-117 U
gi 123728449 gb BU762277.1 BU762277	sar87d02.y1 Gm-c1074 Gly...	421 e-115 U
gi 15811812 gb BE917591.1 BE917591	GmCHS6 soybean root subt...	419 e-114 U

Did you know...
 You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.



NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Go Links

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=23729544&dopt=GenBank

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for [Go] [Clear]

Limits Preview/Index History Clipboard Details

Display GenBank Show: 20 Send to File Features

1: BU762864 sas34a07.y1 Gm-c1. [gi:23729544]

Links

LOCUS BU762864 527 bp mRNA linear EST O2-JUL-2004

DEFINITION sas34a07.y1 Gm-c1080 Glycine max cDNA clone SOYBEAN CLONE ID: Gm-c1080-5414 5' similar to TR:Q9XEX5 Q9XEX5 CYTIDINE DEAMINASE ; mRNA sequence.

ACCESSION BU762864

VERSION BU762864.1 GI:23729544

KEYWORDS EST.

SOURCE Glycine max (soybean)

ORGANISM Glycine max
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids ; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae; Glycine.

REFERENCE 1 (bases 1 to 527)

AUTHORS Shoemaker,R., Keim,P., Vodkin,L., Erpelding,J., Coryell,V., Khanna ,A., Bolla,B., Marra,M., Hillier,L., Kucaba,T., Martin,J., Beck,C., Mylie,T., Underwood,K., Steptoe,N., Theising,B., Allen,M., Bowers ,Y., Person,B., Swaller,T., Gibbons,M., Pape,D., Harvey,N., Schurk ,R., Ritter,E., Kohn,S., Shin,T., Jackson,Y., Cardenas,M., McCann ,R., Waterston,R. and Wilson,R.

TITLE Public Soybean EST Project

JOURNAL Unpublished (1999)

COMMENT Contact: Shoemaker R/Public Soybean EST Project
Public Soybean EST Project

Internet

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Go Links

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=23729544&dopt=GenBank

200401) was used to synthesize the cDNA. First-strand synthesis was performed with 5'-methyl dCTP, hence the ligated cDNA was hemimethylated. A modification of Stratagene's first-strand synthesis primer was used. An 'anchor' nucleotide (V=A, C, or G) was added to the 3' end of the primer [GAGAGACAGAGAGAGAGAACATGCTCGAG(T)18V] to anchor the primer at the 5' end of the poly(A) tract. After second-strand synthesis, the cDNA ends were filled in with cloned Pfu DNA, ligated to EcoRI adaptors and subsequently phosphorylated. The cDNA was then precipitated and redissolved in sterile, RNase-, DNase-free water. The XbaI site within the first-strand synthesis primer was then restricted by digestion with XbaI from Promega (40U/ul); all XbaI sites in the cDNA would be protected by their hemimethylated status. The cDNA constructs were size-fractionated with a 500bp cutoff, using Sephadryl S-500 High Resolution (Pharmacia Biotech) in a 2-mm diameter column and a bed volume of approximately 1ml. The column eluent was precipitated, redissolved, and ligated into Stratagene's pBluescript II XR Predigested vector (pBluescript II SK(+)) vector that has been digested with EcoRI and XbaI, and phosphorylated by Stratagene. This library was constructed in the laboratory of Dr. Paul Keim and Dr. Virginia H. Coryell at Northern Arizona University."

ORIGIN

```
1 tacagcgcgt cccctcccg cgtcgcgtt ctgcattcca agggaaatgt ttttaaaggc
61 ttctacatgt agtcgcgtgc ttataacccc agcttggggc cgcttcaggc cgccatcgct
121 gccttcatcg cggccggccg tggggattat gaagagatgt ttggccgggt gttggggag
181 aagaagggg cggtcatcaa acaggatcac actgcgaagggt tgctgtctca ttccatcagec
241 ccacgtgcc acttcaaacaa ttttttgtgttgcattacttccatctt aacattgttt
301 tttttttctt etaccatcaa ttaataataat atataacta ttggagatgt gattttgtt
361 cacccatcacc tgacatcacc ttgtataataat tgatgtgtca cggatagaca tggtgtgtt
421 gtatataat ccccttccta ttatgccta tgggttctaa atatttaa acaagcttt
481 ttttttgta tattgttaca tctaataatgtt gatattgtcc tgtcaaa
```

//

Internet

FASTA/BLAST Statistics

- * E() value is equivalent to standard P value
- * Significant if E() < 0.05 (smaller numbers are more significant)
 - The E-value represents the likelihood that the observed alignment is due to chance alone. A value of 1 indicates that an alignment this good would happen by chance with any random sequence searched against this database.
- * The histogram should follow expectations (asterisks) except for hits

Interpretation of output

- * very low E() values ($e-100$) are homologs or identical genes
- * moderate E() values are related genes
- * long list of gradually declining of E() values indicates a large gene family
- * long regions of moderate similarity are more significant than short regions of high identity

What this does for you

- ⌘ You identified what gene is encoded by your clone's sequence
- ⌘ Perhaps you may have found the function of your gene
- ⌘ You have more cDNA sequences to add together to build a consensus and perhaps a full-length cDNA

Biological Relevance

- ⌘ It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- ⌘ Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- ⌘ Are the mismatches conservative ones?
- ⌘ Are the matching regions important structural components of the genes or just introns and flanking regions?

Borderline similarity

- ★ What to do with matches with E() values in the 0.5 -1.0 range?
- ★ this is the “**Twilight Zone**”
- ★ retest these sequences and look for related hits (not just your original query sequence)
- ★ similarity is transitive:
if **A~B** and **B~C**, then **A~C**

Advanced Similarity Techniques

Automated ways of using the results of one search to initiate multiple searches

- ★ **INCA (Iterative Neighborhood Cluster Analysis)**
<http://itsa.ucsf.edu/~gram/home/inca/>
 - Takes results of one BLAST search, does new searches with each one, then combines all results into a single list
 - JAVA applet, compatibility problems on some computers
- ★ **PSI BLAST**
<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html>
 - Creates a “position specific scoring matrix” from the results of one BLAST search
 - Uses this matrix to do another search
 - builds a family of related sequences
 - can’t trust the resulting e-values

File Edit View Favorites Tools Help

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=14008724&dopt=GenBank

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for [Go] [Clear]

Limits Preview/Index History Clipboard Details

Display GenBank Show: 20 Send to File Features

1. BG725328 sae35d12.y1 Gm-c1...[gi|14008724]

LOCUS BG725328 510 bp mRNA linear EST 22-JUL-2004

DEFINITION sae35d12.y1 Gm-c1051 Glycine max cDNA clone GENOME SYSTEMS CLONE

ID: Gm-c1051-7103 5' similar to SW:CHS6_SOYBN P30080 CHALCONE SYNTHASE 6 ; mRNA sequence.

ACCESSION BG725328

VERSION BG725328.1 GI:14008724

KEYWORDS EST.

SOURCE Glycine max (soybean)

ORGANISM Glycine max

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids ; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae; Glycine.

REFERENCE 1 (bases 1 to 510)

AUTHORS Shoemaker,R., Keim,P., Vodkin,L., Erpelding,J., Coryell,V., Khanna ,A., Bolla,B., Marra,M., Hillier,L., Kucaba,T., Martin,J., Beck,C., Wylie,T., Underwood,K., Steptoe,M., Theising,B., Allen,M., Bowers ,Y., Person,B., Swaller,T., Gibbons,M., Pape,D., Harvey,N., Schurk ,R., Ritter,E., Kohn,S., Shin,T., Jackson,Y., Cardenas,M., McCann ,R., Waterston,R. and Wilson,R.

TITLE Public Soybean EST Project

JOURNAL Unpublished (1999)

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

File Edit View Favorites Tools Help

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=14008724&dopt=GenBank

```
/db_xref="taxon:3847"
/clone="GENOME SYSTEMS CLONE ID: Gm-c1051-7103"
/tissue_type="floral meristematic mRNA"
/lab_host="DH10B"
/clone_lib="Gm-c1051"
/note="Vector: pBluescript II SK+; Site_1: EcoRI; Site_2: XbaI; The cDNA library was constructed from floral meristematic mRNA provided by Dr. Halina Knap of Clemson University. Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a XbaI restriction site. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by XbaI digestion. The cDNA fragments were directionally cloned into the EcoRI-XbaI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into DH10B host cells (GibcoBRL). This library was constructed in the laboratory of Dr. Randy Shoemaker."
```

ORIGIN

```

1 ggttgtgcac agtccatttt ctattggactt cttcctcatt tggatccaaga tgaacagcac
61 acatgcacctt cctaacttag ctcaacttg aacgacgtgt cttagatgt ccatttttc
121 atgcgttcat cctaacttag ctcaacttg gtcaaaaat gctggccac cagggtgtgc
181 aatccaaatg atagagtgtt aatcatcaat ttcaatgggt ttgaagggtt caaccaaggc
241 cttttcgatg ttcttggaga tgatgtccagg aacatccctg aggagatggaa aagtgtgtc
301 tacttggccc aggtggccat caaatggccc ttctgtgtttt ggaaggatgt ttttgtcgat
361 ccacacaaggc tcaaacaaatg ttcttccatgg tggccagagga tctgtatccaa caatgcacgc
421 agtcgtccacc ttctccaaatc aggtttgtcccc cacaaaggctg tcaatgttgc tggtaactccg
481 gccacgaaat gtgcactgtgt tgatctcccg
//
```

[Disclaimer](#) | [Write to the Help Desk](#)
[NCBI](#) | [NLM](#) | [NIH](#)

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

FASTA format

- * One of three formats used for sequences
- * Begins with single-line description followed by sequence data
- * Description line starts with ">"
- * Example:

```
* >gi|532319|pir|TVFV2E|TVFV2E envelope protein  
ELRLRYCAPAGFALLCNDADYDGFKTNCSNVVHCTNLMNTTGTLLNGSYENRT  
QIWKHRTSNDALILLNKHYNLTVCKRPGNKTLPVTIMAGLVFHSQLRLRQAWC  
HFPSNWKGAWKEVKEEVNLPKERYRGTDPKRIFFQRQWGDPEANLWFNCHGEFFYCK  
MDWFLNYLNLLTVDAHDNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK  
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVLSPQIESIWAELDRYKLVEITPIGF  
APTEVRRTGGHERQKRVPFXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL  
LAAVEAQQQMLKLTIWGVK
```

The nucleic acid codes supported are:

A → adenosine	M → A C (amino)
C → cytidine	S → G C (strong)
G → guanine	W → A T (weak)
T → thymidine	B → G T C
U → uridine	D → G A T
R → G A (purine)	H → A C T
Y → T C (pyrimidine)	V → G C A
K → G T (keto)	N → A G C T (any)

- gap of indeterminate length

accepted amino acid codes are:

A alanine	P proline
B aspartate or asparagine	Q glutamine
C cystine	R arginine
D aspartate	S serine
E glutamate	T threonine
F phenylalanine	U selenocysteine
G glycine	V valine
H histidine	W tryptophan
I isoleucine	Y tyrosine
K lysine	Z glutamate or glutamine
L leucine	X any
M methionine	* translation stop
N asparagine	

- gap of indeterminate length

Types of data integrated in genome browsers

- Genomic sequence
- RefSeq mRNAs (non-redundant)
- GenBank mRNAs (redundant)
- ESTs
- Gene predictions
- SNPs
- Homologous sequences from other organisms
- STSs

Other Sequence Search Tools

- SRS (Sequence Retrieval Service) was created by Dr. Thure Etzold: [*CABIOS* 9(1); 49-57 (1993)]
- It is a meta search engine for all types of biological data in hundreds of databases as well as about 20 sequence analysis programs
- SRS can be accessed over the WWW from many servers (mostly in Europe):

<http://srs.ebi.ac.uk/>

<http://www.infobiogen.fr/srs6bin/cgi-bin/wgetz?-page+top>

<http://www.sanger.ac.uk/srs6bin/cgi-bin/wgetz?-page+top>

<http://iubio.bio.indiana.edu/srs6bin/cgi-bin/wgetz?-page+top>



Netscape: Query Form

Back Forward Reload Home Search Netscape Images Print Security Shop Stop Location: http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz What's Related

TOP PAGE QUERY RESULTS SESSIONS VIEWS DATABASES HELP

search EMBL EMBLNEW SWALL SWISSPROT SPTRREMBL REMTRREMBL TREMBLNEW ENSEMBL PATENT PRTJPO PRTPATENT DNAUSPO PRTIMGT IMGTHLA

Reset Info about field AllText

Submit Query

append wildcards to words

combine searches with AND

Number of entries to display per page 30

retrieve entries of type Entry

Use predefined view * Complete entries *

Extended query form

Select fields to display:

ID
AccNumber
Description
Keywords
Organism
SeqLength
Feature: ID

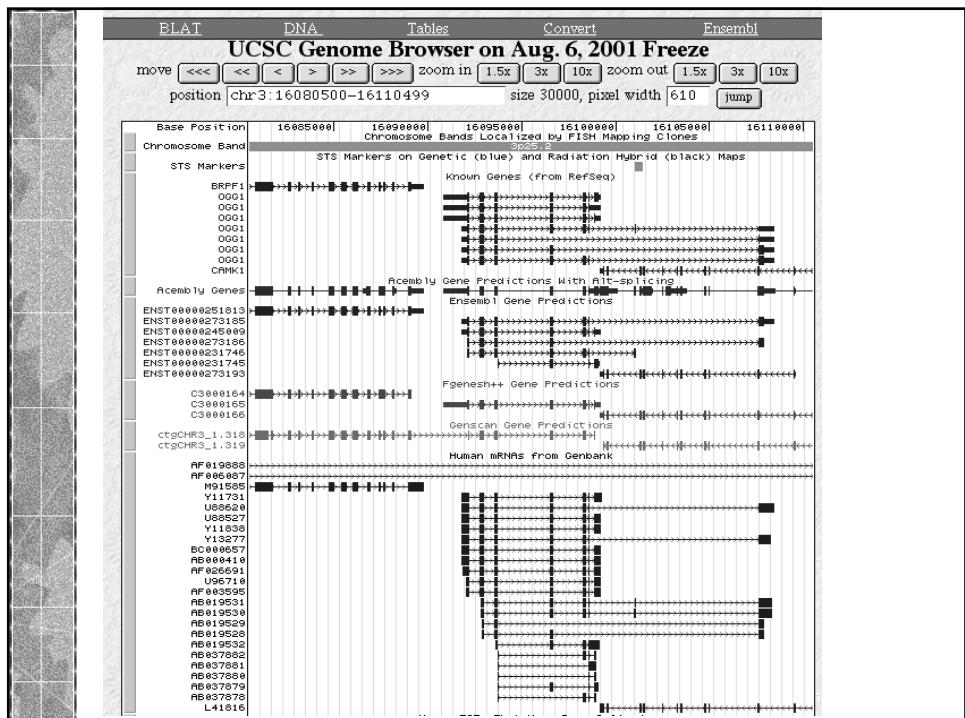
sequence format genbank

Why So Many Databases?

- ★ If GenBank has all sequence data and Entrez is such a good query tool, then why are there so many other sequence databases?
 - Specialized data (single species, immunoglobulins, etc.)
 - Better annotation (i.e. SwissProt)
 - Sequences linked to other data (ACEDB)
 - Subornness and local pride - EMBL, DDBJ
- ★ Well designed databases are interlinked with others for supplemental data
- ★ It is very hard to get all relevant information across all databases for any gene

Other Genetic Databases

- * Genome Sequence - where does a gene fall on the genome
 - integrate multiple layers of information
 - > Sequence contigs, mRNAs, predicted exons, etc.
 - Single species?
- * ESTs: dbEST @ NCBI
- * SNPs: dbSNP @ NCBI,
<http://snp.cshl.org> (SNP Consortium)
- * Metabolism/Pathways
- * Gene Function (Genome Ontology)
- * Protein motifs/domains and protein families



Genome Databases

- * New area - in desperate need of development
 - Chromosomes::Sequence::Contigs::Clones::
 - STS Markers::Genetic Markers::Genes::
 - Features::Expression data::Phenotype
- * No single database can hold it all
- * UCSC is probably the best right now
genome.ucsc.edu
- * Need a data exchange and linkage infrastructure

European Bioinformatics Institute

- * Products and services
- * Databases
 - ◆ Literature
 - ◆ Microarray
 - ◆ nucleotide
- * Toolbox with software
 - ◆ Similarity searches
 - ◆ Protein function
 - ◆ Sequence analysis
 - ◆ Structure
- * <http://www.ebi.ac.uk/>

EBI Services - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Favorites Go Links

Address http://www.ebi.ac.uk/services/

EBI Home About EBI Research Services Toolbox Databases Downloads Submissions

VIEW ALL EBI SERVICES

Services Overview

Databases

- EMBL via WEBIN
- EMBDep
- IMGT/HLA
- PDB-AutoDep
- UniProt via SPIN
- Webin-Align

Toolbox

Similarity & Homology

- Blast2 - ASD IIEW
- Blast2 - EVEC
- Blast2 - NCBI
- Blast2 - Parasite
- Blast2 - VJU
- Fasta
- Fasta - LOC IIEW
- Fasta - GenoProt
- MPsrch
- more...

Prot. Function. Analysis

- CluStr
- GeneQuiz
- InterProScan
- more...

Sequence Analysis

- Align
- ClustalW Updated
- GeneMse
- PromoterMse
- more...

Structural Analysis

- DALI
- Maxsprout

Databases

WHAT'S 2can?
This logo is a link to a relevant section in the EBI's new bioinformatics educational website, '2can Bioinformatics'.

Literature Databases

- MEDLINE
- MIM
- Patent Abstracts
- more...

Microarray Databases

- ArrayExpress
- MAME

Nucleotide Databases

- ASD
- EMBL-Bank
- Ensembl
- Genome Reviews
- IMGT/HLA
- more...

eDualHead Toolbar L07647 Glycine max ... Document11 - Microsoft Word W11_12_Summary h_Phylogenetic Trees b_DNA retr

IDENTIFIERS

dbEST Id: 101883
EST name: yb01a01.s1
GenBank Acc: T48601 **GenBank gi:** 650461
GDB Id: 490761

CLONE INFO

Clone Id: IMAGE:69864 (3')
Other ESTs on clone: yb01a01.rl
DNA type: cDNA

PRIMERS Sequencing: -21m13
PolyA Tail: Unknown

SEQUENCE

```
GGCGGCTCAGTAGCAGGTGCCGTCCACCTCCGCCATGACAACAGACATGACATGGT
GGTTTACACCAAGCGTCGATGCTCTCTGTGAAGGGCAGCCAGGGCCTCCATTG
CACCATCGAGGAGAAGGNTCCCCCTTCTCCAGSTCTCGGTGCCACCGCAGATATGCT
GGTCACAGAAGGTGTTGGTGCCTGGTGGNTCTNCANGATGCCAAGTCAGGTACT
TNTGGGGCAGCTTGACGGCTTCAACCGGTCNNCCAGCTTCTTCAGGGCCANCTTC
AACCNNGGCTACAGGCCTTAACCGGTTCAACCGGTCNNCCAGCTTCTTCAGGGCCANCTTC
```

Quality: High quality sequence stops at base: 277
Entry Created: Feb 6 1995 **Last Updated:** Feb 6 1995

COMMENTS

High quality sequence stops: 277
Source: IMAGE Consortium, LLNL
This clone is available royalty-free through LLNL ; contact the IMAGE Consortium (info@image.llnl.gov) for further information.

PUTATIVE ID Assigned by submitter
similar to gb:S71043_rnai IG ALPHA-2 CHAIN C REGION (HUMAN)

LIBRARY

Lib Name: Stratagene placenta (#937225)
Organism: Homo sapiens
Sex: male
Organ: placenta
Lab host: SOLR cells (kanamycin resistant)
Vector: pBluescript SK-
R. Site 1: EcoRI
R. Site 2: XbaI
Description: Cloned unidirectionally. Primer: Oligo dT. Caucasian.
Average insert size: 1.2 kb; Uni-ZAP XR Vector; ~5' adaptor sequence: 5' GAATTCGGCACAGAG 3' ~3' adaptor sequence: 5' CTCGAGTTTTTTTTTTTTTTT 3'

Database Search Strategies

- ⌘ General search principles - not limited to sequence (or to biology)
- ⌘ Use accession numbers whenever possible
- ⌘ Start with broad keywords and narrow the search using more specific terms
- ⌘ Try variants of spelling, numbers, etc.
- ⌘ Search all relevant databases
- ⌘ Be persistent!!**

What we covered today

- ⌘ Retrieving a known DNA sequence
- ⌘ Similarity searching with a DNA sequence
- ⌘ BLAST